BIG PROBLEMS WITH BIG DATA
The Case for Right Data v/s Big Data

April 24, 2019

Written by
Pavan Bagai
President and Chief Operating Officer

lookdeeper@exlservice.com
“I have a very good idea of the problem I need to solve. The catch is that my data is inadequate, widely dispersed and worst of all, incomplete. Can you still build me a viable model?”

In those days, the first step in most analytics projects was to compile and cleanse available data and deploy advanced techniques to fill the gaps without biasing its integrity.

Today most clients are in a diametrically opposite position:

“I have reams of all sorts of data that increases by the minute. I am not sure what it contains but I believe it must have some value. Can you dig down and sift through it and conjure up some meaningful insights that may solve some business problem for me?”

This 180-degree reversal is an outcome of the explosive growth of Big Data. This phenomenon has been fed by two main factors:

1. An exponential increase in the generation of data in real time from a variety of digital sources – internet activity, smart phones, credit card machines, ATMs, GPS devices, installed and wearable sensors, IoT and a host of others. This data relates to transactions, locational mobility, usage, activity, behavior and opinions at the individual level,

2. A rapid expansion in computational power and digital data storage capacity, accompanied by development and adoption of advanced data processing and statistical techniques.

Big Data has catalyzed the creation of open source software frameworks, advances in analytical methods such as Machine and Deep Learning, and impressive progress in the field of Artificial Intelligence. It has also spawned a corps of data scientists and Chief Data Officers (CDOs), sensational headlines and articles in magazines, talking material for Futurists and a generous dose of buzz and hype.

There is a popular perception that Big Data algorithms can work miracles by generating better insights, making more accurate predictions and feeding AI applications. This is not necessarily true in all cases.

It is useful to classify data along two dimensions – Structured v/s Unstructured and Internal v/s External.

When plotted graphically, data in the resultant four quadrants possess significantly different characteristics. Also, the quadrants are not of uniform size which reflects substantial variations in volume of data in each category.

CDOs often refer to their data repositories as “data lakes.” We will extend this water body metaphor to characterize the four types of data.
We will now subjectively assess and rate each category on five relevant criteria.

The purpose of this classification scheme is not to rank the four categories of data in order of overall superiority. It is to better understand the pros and cons of each type and the effort required to extract utility from it. Different business situations and different strategies will require combinations of these data types in varying proportions.

It should also be clear that there is a very high degree of “noise” in Big Data. In most cases we do not fully know what it contains and not all of it is useful. Filtering out the cacophony to be able to listen to the melody within is akin to sieving through a murky swamp to find a small gold nugget. While searching for value in this morass is a worthy objective, a large portion of the effort can be wasteful.

Boasts about how many tera-, peta- and exabytes of data we generate or store are hollow. What matters is what miniscule fraction of it is of value. How “Big” the data is, is not important. What is important is whether it is the “Right” data.
Interestingly, Human Intelligence, which Artificial Intelligence is expected to mimic, does not operate on a similar basis. Imagine if a child were expected to read 2,000 pages of text to learn how to read in the first place! Quite paradoxical!

Human beings are able to learn concepts, abstractions and associations from a single or handful of observations. They view objects as an amalgam of features, not of pixels. They can certainly enhance their ability with practice and experience but only after the conceptual foundation is initially laid.

Metaphysics and Epistemology may suggest that human intelligence and consciousness have facets that are not fully understood and hence replicating them artificially may be a bridge too far. Such arguments notwithstanding, researchers in the field of Cognitive Science are developing techniques to power Artificial Intelligence that more closely resemble Human Learning.

### Pros and Cons of Structured and Unstructured Data

<table>
<thead>
<tr>
<th></th>
<th>INTERNAL</th>
<th>EXTERNAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STRUCTURED</td>
<td>UNSTRUCTURED</td>
</tr>
<tr>
<td></td>
<td>Man-made Lagoon</td>
<td>FISH POND</td>
</tr>
<tr>
<td>How well is data</td>
<td>How well is data</td>
<td>How well is data</td>
</tr>
<tr>
<td>Organized?</td>
<td>organized?</td>
<td>organized?</td>
</tr>
<tr>
<td>Do we know what it</td>
<td>How expansive is the</td>
<td>How expansive is the</td>
</tr>
<tr>
<td>contains?</td>
<td>content?</td>
<td>content?</td>
</tr>
<tr>
<td>How expansive is the</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>content?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How useful is the</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>content?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How well is the</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>data maintained?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Methods that can generate learning from few observations and are therefore not reliant on vast and wasteful Big Data sets are likely to be significantly more efficient. Similarly, techniques that can promptly discard large volumes of useless information early in the process will also provide greater efficiency. Costs related to generation, acquisition, storage and management of petabytes of data can be avoided. These techniques are based on more intense computations applied to sparse but Right (relevant) data.

Bayesian Program Learning is one such area that is receiving considerable attention. Experiments related to visual recognition have demonstrated this technique to be superior to Machine Learning and, in some cases, even humans.

The next phase of development in Cognitive Science is likely to focus on improvements in efficiency of Big Data methods. While this could take various directions, the two most likely areas are:

1. Development and adoption of techniques that more closely replicate human learning methods and hence don’t require Big Data, but Right Data.

2. Honing of filtration techniques in Data Science such that irrelevant data (“noise”) is eliminated prior to the application of computation cycles. Only Right Data within Big Data is then worked upon.

In either case, quality of data will take precedence over quantity of data. The current obsession with Big Data is likely to be tempered and the benefits of Right Data more widely appreciated.

In practical terms, we are likely to see a marked increase in adoption of Probabilistic Programs and Generative Analytical Models that encode human experience into algorithms at the expense of the data-hungry Machine Learning models of today.

This should lead to diversion of investment from costly and relatively inefficient Big Data management programs to accelerating the development and adoption of techniques that represent the next wave of Artificial Intelligence in business.

Ideally we should see “Right” Data supersede “Big” Data. Perhaps BRight Data may catch on as a buzz term. Then, Artificial Intelligence could well become less “Artificial” and more genuine!

**Acknowledgements**

I would like to thank Ankor Rai, Nagendra Shishodia, Nalin Miglani and Vivek Jetley for critiquing earlier drafts and providing valuable suggestions. Errors that may remain are solely on my account.

–Pavan Bagai
EXL (NASDAQ: EXLS) is a leading operations management and analytics company that designs and enables agile, customer-centric operating models to help clients improve their revenue growth and profitability. Our delivery model provides market-leading business outcomes using EXL’s proprietary Business EXLerator Framework®, cutting-edge analytics, digital transformation and domain expertise. At EXL, we look deeper to help companies improve global operations, enhance data-driven insights, increase customer satisfaction, and manage risk and compliance. EXL serves the insurance, healthcare, banking and financial services, utilities, travel, transportation and logistics industries. Headquartered in New York, New York, EXL has more than 27,000 professionals in locations throughout the United States, Europe, Asia (primarily India and Philippines), South America, Australia and South Africa.

© 2018 ExlService Holdings, Inc. All Rights Reserved. For more information, see www.exlservice.com/legal-disclaimer